LA-UR -91-1512

Conf - 910810

LA-UR--91-1512

DE91 013522

Los Alamos National Laboratory is operated by the University of California for the United States Department of Energy under contract W-7405-ENG-36

TITLE COMPARISON OF FOUR METHODS FOR AGGREGATING JUDGMENTS FROM MULTIPLE EXPERTS

AUTHOR(S): JANE M. BOOKER
RICHARD R. PICARD

SUBMITTED TO PROCEEDING OF JOINT STATISTICAL MEETINGS - STATISTICAL COMPUTING SECTION
ATLANTA, GA
AUGUST 19 - 23, 1991

DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

Los Alamos

Los Alamos National Laboratory
Los Alamos, New Mexico 87545

COMPARISON OF FOUR METHODS FOR AGGREGATING JUDGMENTS FROM MULTIPLE EXPERTS

by
Jane M. Booker and Richard R. Picard
Statistics Group, MS F600
Los Alamos National Laboratory, Los Alamos NM 87545

I. INTRODUCTION

This report describes a study that compares four different methods for aggregating expert judgment data given from multiple experts. These experts need not be a random sample of available experts. The experts estimate the same unknown parameter value (e.g., a failure rate or a probability). Their estimates need not be a representative set of sample values from an underlying distribution whose mean is an unknown parameter, θ. However, it is desired to combine the experts' estimates into a single aggregation estimate to reflect their amount of available knowledge about the unknown parameter.

Many different aggregation estimators and methods have been proposed in the literature (Meyer and Booker, 1991). However, few have been used, tested, or compared. Four different methods are chosen for this study which have been used or proposed for use in NRC studies. The set represents a cross section of the various types of methods.

The results of this study do not indicate the use of any one method over another. Methods requiring minimal decision maker input are sensitive to the biases in the experts' responses. For these methods, there is no mechanism to adjust the experts' estimates to account for any known biases in the expert population such as optimism or pessimism. The results of this study indicate that these methods tend to perform poorly in all but the most ideal cases. Conversely, methods requirnng extensive decision maker inputs are sensitive to misspecification. These methods perform poorly unless complete information is known about all the experts. That is, the decision maker's input parameters must nearly equal the actual values.

This report is divided into eight sections as follows: The four methods are described in the second section. The third section contains a description of the example from NUREG 1150 that is used in the study. The fourth section is a preliminary examination of the first and most complex method. This investigation examines the different parameters and includes an investigation on the effects on the aggregation parameters of the correlation structure. To compare the methods, an underlying set of actual parameter values was established. The set is described in the fifth section. The sixth and seventh sections describes the design of the computational runs made in the sensitivity study and in the simulations. The sensitivity study consists of a set of calculations that compares three of the methods run on similar cases. Two simulations runs are designed to investigate the second and fourth methods in separate analyses. Section VII ends with a description of the measures used to compare the methods with a set of actual values. The results of the runs are given in Section VIII where the performance of each method is evaluated using the measures of comparison relative to the actual values established. The final section includes the conclusions drawn from the results and a discussion of recommended uses of the methods. The Appendix contains the notation used in the report.

II. DESCRIPTIONS OF THE FOUR METHODS

The Lindley and Singpurwalla (LS) Method

Lindley and Singpurwalla (1986) proposed a method for aggregating expert estimates based on normal theory. The particular estimates they cite are component failure rates, Λ, having lognormal distributions. Therefore, the quantity, $\theta = \ln(\Lambda)$, is normally distributed. Each expert (i) states his estimates for the mean (m_i) and standard deviation (s_i) of his subjective underlying (normal) distribution of θ. They assume that there is a decision maker who will aggregate these means and variances according to his knowledge of the experts. He must have sufficient information to estimate four different parameters for each expert such that the expert's stated mean (standard deviation) will be adjusted (weighted) in a predetermined direction. This direction is determined by the decision maker, and there are no guarantees that this determination is in the correct direction (toward the actual mean).

The parameters α, β, γ, and ρ represent a location adjustment parameter for the expert's mean, a scaling parameter for the expert's mean, a scaling parameter for the expert's standard deviation, and the correlation among the experts, respectively. Experts are subject to many different kinds of biases that affect their estimates. The scaling and location adjustment parameters can be used to counter some of these biases (Kahneman, Slovic, and Tversky, 1982). Location adjustments are useful to counter the human tendency to underestimate probabilities of uncommon events, to counter optimism or pessimism, and to counter the tendency to overestimate the occurrence of extremely rare events. Scaling adjustments are useful to counter the tendency of humans to underestimate uncertainty and variability (Lindley and Singpurwalla, 1986).

The decision maker's final estimate of θ is normal with mean

$$\mu = \sum_{i,j} \beta_i \sigma^{ij}(m_j - \alpha_j) / \sum_{i,j} \beta_i \sigma^{ij} \beta_j$$

and standard deviation of that mean

$$\sigma = \left[\sum_{i,j} \beta_i \sigma^{ij} \beta_j \right]^{-1/2}$$

where σ^{ij} are the elements of the matrix Σ^{-1} such that the matrix Σ has off diagonal elements $\sigma_{ij} = \rho_{ij} (\sigma_{ii}\, \sigma_{jj})^{1/2}$ and diagonal elements $\sigma_{ii} = (\gamma_i\, s_i)^2$. This result is developed and labeled Theorem 1 in their paper.

The final estimate is in the general form of a weighted mean. The other parameters combine as weights for the experts' mean estimates. This structure is important later on in comparing the LS method to the weighted mean method.

Three assumptions are required for this theorem. First, the experts' estimates of the standard deviation, (s_i), must provide no information on θ. Second, the decision maker's prior distribution on θ is effectively constant, implying that the decision maker's knowledge of θ is weak prior to viewing the experts' estimates. Third, the distribution of the experts' means, conditioned on their stated standard deviations, is a multivariate normal with each mean, m_j, having a mean $\beta_j\theta + \alpha_j$, standard deviation of $\sigma_j s_j$, and correlation ρ_{ij} between m_i and m_j. In addition, because the correlations and the γ values are determined by the decision maker prior to or without knowledge of the experts' estimates, the Σ matrix may not be invertible. This method requires that it be invertible.

It may be very difficult to satisfy the normality assumptions in applying this method to real problems. Very often what little data exists (including expert estimates) does not support normality (or lognormality). Expert estimates often indicate multimodal distributions (Booker and Meyer, 1988), and it can be difficult to justify any distribution type when there are five data points or less . Lindley and Singpurwalla cite one of the few cases where a lognormal distribution may apply. It should be noted that the simulated data chosen for this study follows the correct distribution assumption, therefore, providing a best case for the LS method. The effect of non normality may or may not be severe drawback. Sensitivity to this assumption is not examined in this study. Other features of the method regarding its use are discussed below.

This method addresses the issues of correlation among experts and characterizing uncertainty in the experts' estimates. Many methods, including some used in this study, do not consider the correlation issue. One reason for that is because correlations among experts are difficult (or impossible) to estimate very well. Such estimation requires an extensive and controlled study of the experts, how they solve problems, and how they interact. Such studies are usually not practical. On the other hand, most recent methods incorporate some estimation or handling of uncertainty. In the LS method, uncertainty in the experts' estimates is characterized by the variance, s^2, provided by the experts.

However, this method is difficult to use because of detailed input required from the decision maker. In speaking with Singpurwalla, we found that he was unaware of any actual use to date. In addition, a literature search indicated five citations of the LS paper but no examples of actual use. This was discouraging because there was no guidance on how decision makers should estimate the four sets of parameters. In our experience (Meyer and Booker, 1991), decision makers do not have the knowledge required to accurately estimate these parameters. In using this method, they are forced to provide a set of values for all the parameters using little or no information. In addition, the results of this study will indicate, conclusions can be very sensitive to the decision maker's parameters.

The Self Weights (SW) Method

The second method was previously used in applications involving aggregating expert estimates to determine probabilities and characteristics of seismic events (Bernreuter, et al., 1989). The aggregation estimator is a weighted mean of the estimates given by the experts. The weights are determined by the experts themselves and represent the experts' evaluations of their own levels of expertise. It is not apparent from the report whether or not this self evaluation is done considering the expert's level relative to colleagues or relative to some absolute scale. At any rate, the weights are chosen from a numerical scale of integers, 1-10, where 10 is the highest level of expertise. The analyst then normalizes the weights for the experts.

The report did not specify any corresponding variance for this weighted mean. Consequently, the most basic question of interest (whether the resulting interval formed from the aggregated mean and aggregated uncertainties covers θ) cannot be evaluated. In the section below on modifying the methods for the simulation study, the variance for the weighted mean is found by using the statistical theory.

This method lacks several features of the first method. There is no variance specified for the aggregation estimator, there is no mechanism for handling expert correlation, and there is no provision made for the experts to provide uncertainties on their estimates.

Unlike the first method, this method makes no assumptions on the data, and the analyst does not have to estimate any parameters to do the aggregation.

It is easy to criticize the use of the self weights in this method. What little evidence exists supports the cognitive theory that experts cannot evaluate themselves (or even others) very well or consistently and that different experts' ratings do not agree very well

(Kahneman, et al., 1982). Even one of the report authors admitted that asking experts for self weights was not a good idea (Mensing, 1990).

The Equal Weights (EW) Method

The third method was taken from two reports (McCann, et al., 1988; McGuire, et al., 1989) by the Electric Power Research Institute (EPRI), which used expert estimates to characterize seismic events and probabilities. In the first report, a weighted mean was chosen for the aggregation estimator. Because the authors felt that the weights were chosen in an arbitrary manner, the second report recommended weighting all experts the same. Therefore, this method is used as an equal weights method.

The standard deviation used in the report is a composite formula representing two sources of variation. The first is an average of the variances provided by the experts. The second is the variation of among experts' estimates. The total variance is the sum of these two sources

$$\sigma^2 = \frac{1}{n}\sum_i s_i^2 + \frac{1}{n-1}\sum_i \left(m_i - (\sum_i m_i/n)\right)^2 .$$

This method lacks some of the features of the first method. There is no mechanism for handling expert correlation. No location or scale adjustments of the experts' estimates (similar to the LS method) are possible. The experts provide estimates of uncertainty on their mean estimates through the values of s_i^2. However, this method does not require assumptions on the data, and the analyst does not have to estimate any parameters to do the aggregation.

Equal weights is a commonly used method (Seaver, 1978; Winkler, 1986). The rationalization for its use stems from the usual state of insufficient knowledge necessary to specify unequal weights. Because this method has had some usage, an equally weighted mean is a good candidate for inclusion in this study.

The Method of Empirical Distributions (ED)

A final method was added to the study that was based on aggregating distributions provided by the experts rather than aggregating single estimates. The distributions are elicited from the experts themselves and represent the uncertainty that the experts have concerning the quantity being estimated. The elicited values can be ranges, percentiles, means, or any set of values that can be translated into a cumulative empirical probability distribution.

One application of this type of elicitation is found in the NUREG 1150 study (US NRC, 1989), and this was the method used on the example (Section III) chosen for this study. Other advocates of this type of approach are Meyer and Booker (1991).

An advantage of this approach is that an entire distribution is obtained that represents uncertainty without assuming a particular distributional form and without relying on variance estimates. Studies have shown that experts are more comfortable and more adept at estimating ranges and percentiles than they are at variances (Meyer and Booker, 1991). The analyst uses simulation techniques to combine the distributions, avoiding the difficulties of estimating parameters. However, if the distributions are combined using a weighting scheme, the analyst must determine what weights to use. Another choice that the analyst makes is the type(s) of aggregation estimates to use. The mean, median, variance, and percentiles of the aggregation distribution can be used to summarize the data. Some analysts prefer the median as an aggregation estimator because experts often estimate a median rather than a mean of their distribution of values (Kahneman, et al., 1982). This method allows the use of the median as an aggregation estimator. Therefore, this method is relatively easy to use for both the experts and the analyst, it requires no assumptions on the data, it provides different types of aggregation estimates, and it characterizes uncertainty.

A disadvantage is that no mechanism is readily available to account for correlation among experts or any location and scale adjustments on their estimates. However, it is possible to include calculations in the aggregation process (in the simulation) to include correlation. Another disadvantage is that the various values (percentiles, ranges, etc.) elicited from the experts must have the appropriate interpretation. If an expert is unfamiliar with percentiles and is asked to give 5% tile and 95% tile value, he may be giving a range of values that represents only a 30% or 40% interval. Extreme care in the elicitation process is required to ensure that the expert is properly constructing his empirical distribution of values.

III. THE EXAMPLE USED IN THE STUDY

The NUREG 1150 study elicited many data sets from multiple experts regarding many physical parameters and events. Because the Lindley and Singpurwalla (LS) method was the most complex method, an example data set was sought that fit into its structure. Simplified structures corresponding to the other methods can be viewed as "subsets" of a more complete specification. An example was needed that estimated a single target quantity, preferably a failure rate, that was distributed normally or lognormally. An example that used information from several experts was desired. Also, the experts had to provide some estimates of uncertainty on their failure rates either in the form of distributions, or percentiles, or standard deviations. In Appendix C, Volume 2 of NUREG/CR4550, (Wheeler, et al., 1986), the failure rate for a specific valve rupture scenario at the Sequoyah plant was estimated by five experts. The experts provided cumulative distribution functions and percentile values for the log (base 10) of this failure rate, Λ.

Two conversions were necessary on this raw data to make it applicable to the LS method. First, two experts provided estimates in terms of failures/year, and the other three used failures/hour. All were converted to failures/hour. Second, the raw data was in the form of $\log_{10} \Lambda$. The LS method uses $\ln(\Lambda)$. Therefore, the data was translated to this new logarithm base. Table I gives some percentiles and their translations.

TABLE I
PERCENTILE ESTIMATES FROM EXPERTS' DISTRIBUTIONS

			Expert		
Percentile	A	B	C	D	E
0		-27.6	-25.2	-29.8	-14.4
1				-25.9	
5		-25.3		-25.2	-21.8
10	-22.3				
50	-20.5				-19.2
80			-22.9		
90	-19.1				
95			-20.9		-16.6
99	-17.7	-20.7	-17.8	-20.6	
100		-18.4			-24.1

Using these percentiles and normal theory, the mean and standard deviation were estimated as given in Table II.

TABLE II
EXPERTS' ESTIMATES FROM THEIR DISTRIBUTIONS

expert	mean	standard deviation
A	-20.5	1.3
B	-22.5	1.4
C	-23.8	1.4
D	-22.9	0.7
E	-19.2	1.6

A graph of the experts' distributions is given in Figure 1. The values used in this study are based upon the values in Tables I and II. It is not uncommon to find that the experts' means are significantly far apart from each other, when gauged by their stated standard deviations. One possible explanation of this effect is the commonly encountered human bias of underestimating uncertainty (Kahneman, et al., 1982) which appears in misspecification of percentiles (an estimated 90th percentile value should have been a 70th percentile). Another possible explanation is that the large separations in the means are due to genuine differences in opinions.

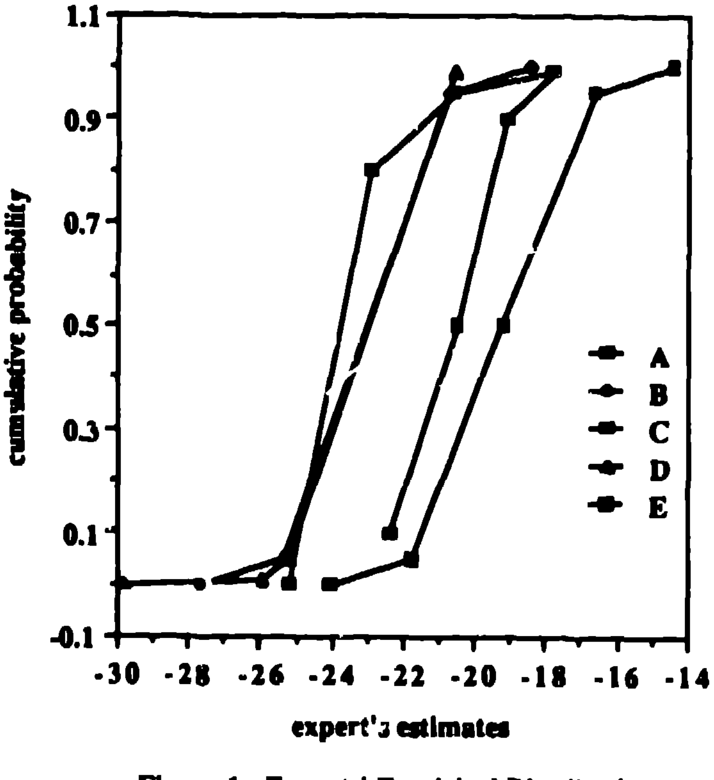

Figure 1: Experts' Empirical Distributions

IV. PRELIMINARY INVESTIGATION OF THE LS METHOD

Investigating the Parameters

Because the LS method has the most complex parameter structure, an initial investigation was made to determine the general effects on the aggregation mean and variance by changes in the parameters. It is noted that using the mean and standard deviation formulas (from Theorem 1) produces the following general relationships:

1) If all αs are identical, increasing α decreases μ (the aggregation mean) by that amount, but do not affect σ.
2) Changes in β rescale both μ and σ by that amount. If all βs are identical, doubling β halves μ, and halves σ.
3) Changes in γ rescales σ, but does not affect μ . If all γs are identical, doubling γ doubles σ.

If the set $\{\alpha_i, \beta_i, \gamma_i\}$ is not the same for all experts and different parameter changes are made for each expert, the relationships are not so straightforward.

The effects on μ and σ by changing the correlation structure were not readily apparent from their formulas or from calculations. A more in-depth investigation follows.

Investigating the Correlation Structure

Because the LS method contains a unique feature specifying the correlation structure among experts, a separate investigation attempted to understand the effect of this structure on the standard deviation of the aggregation mean.

A simplified structure of the LS correlation matrix, P, based on a common correlation ρ among all experts, has the following form

$$P = \rho\, \underline{1}\, \underline{1}^t + (1 - \rho)\, I$$

where superscript ^{t} is the vector transposition, $\underline{1}$ is the vector of all 1s, I is the identity matrix, and the off-diagonal matrix elements are all equal to a common value ρ. This common value of ρ is a special case for this investigation. In general, all ρ_{ij}s may not be the same.

Some initial calculations were made using the LS method where the correlation matrix was changed while the other parameters remained fixed as:

all $\alpha = 0.0$,
all $\beta = 1.0$,
all $\gamma = 1.0$,
all expert standard deviation estimates, $s = 1.0$,
and expert mean estimates, $\underline{m} = \{-20.5,\ -22.5,\ -23.8,\ -22,9,\ -19.2\}$ (Table II).

These parameter values correspond to the case where the experts' mean estimates are not adjusted for any location or scale biases, and their standard deviation estimates are not rescaled.

The LS standard deviation values, for this case, increased as the value of ρ increased as indicated in Table III. This pattern can be explained using basic variance rules where an increasing positive correlation among variables increases their combined variance.

TABLE III
CORRELATIONS AND LS STANDARD DEVIATIONS FOR CONSTANT EXPERT VARIANCES

ρ	σ
0.00	0.45
0.10	0.53
0.25	0.63
0.50	0.77
0.75	0.89
0.90	0.96

For the case where the expert standard deviations are not all equal, the LS standard deviation need not be a monotonic function of ρ. To most decision makers this phenomena is counterintuitive. For the case in Table IV, using the standard deviations from the NUREG 1150 example in Table II, $\underline{s} = \{1.3,\ 1.4,\ 1.4,\ 0.7,\ 1.6\}$, the LS standard deviation increases for increasing values of ρ until $\rho = 0.5$ and then it decreases.

TABLE IV
CORRELATIONS AND LS STANDARD DEVIATIONS FOR DIFFERING EXPERT VARIANCES

ρ	σ
0.00	0.49
0.10	0.57
0.20	0.62
0.40	0.69
0.50	0.70
0.60	0.69
0.80	0.58
0.90	0.45
0.95	0.33
0.99	0.15

The reason for this changing relationship can be found in the formula for the variance in the LS method:

$$\sigma^2 = (\underline{\beta}^t \Sigma^{-1} \underline{\beta})^{-1}$$

where the (i,j)th element of Σ is $\rho_{ij}\, \gamma_i\, s_i\, \gamma_j\, s_j$. Following the structure of the correlation matrix above, Σ can be expressed as

$$\Sigma = \rho\ (\underline{\gamma}^* \underline{s})\ (\underline{\gamma}^* \underline{s})^t + (1 - \rho)\, D^2$$

where $\underline{\gamma}^* \underline{s}$ is the Hadamard product of $\underline{\gamma}$ and $\underline{s}$ and D is a diagonal matrix whose elements are $(\underline{\gamma}^* \underline{s})_i$. The inverse of Σ has this same form,

$$\Sigma^{-1} = (1-\rho)^{-1} \{D^{-2} - \rho/[1+(n-1)\rho\ ]\ D^{-2}\ (\underline{\gamma}^{*}\underline{s})\ (\underline{\gamma}^{*}\underline{s})^{t} D^{-2}\}.$$

Thus

$$\underline{\beta}^{t}\Sigma^{-1}\underline{\beta} = \frac{1}{1-\rho}\sum_{i}\left(\frac{\beta_i}{\gamma_i s_i}\right)^2 - \frac{\rho}{(1-\rho)[1+(n-1)\rho]}\left(\sum_{i}\frac{\beta_i}{\gamma_i s_i}\right)^2$$

which can be rewritten using $\delta_i = \beta_i / \gamma_i s_i$ as

$$\underline{\beta}^{t}\Sigma^{-1}\underline{\beta} = \frac{1}{1-\rho}\sum_{i}(\delta_i-\bar{\delta})^2 + \frac{n}{1+(n-1)\rho}\,\bar{\delta}^2 > 0$$

where

$$\bar{\delta} = \sum_{i} \delta_i/n$$

and n is the number of experts. If $\rho < -1/(n-1)$, then the correlation matrix is not positive definite. It is possible for experts to be negatively correlated. Not much attention is paid to this posibility because negative correlation is difficult to explain and interpret. A decision maker could mistakenly construct a correlation structure which is not positive definite.

Assuming the variance matrix is positive definite, then differentiating the inverse of the variance with respect to ρ gives

$$\frac{\partial\underline{\beta}^{t}\Sigma^{-1}\underline{\beta}}{\partial\rho} = -\frac{(n-1)}{(1-\rho)^2}S_{\delta}^2 + \frac{n(n-1)}{\{1+(n-1)\rho\}^2}\bar{\delta}^2$$

where

$$S_{\delta}^2 = \sum_{i}(\delta_i-\bar{\delta})^2/(n-1) \quad ,$$

which is the "sample variance" of the $\{\delta_i\}$. The differentiation above corresponds to finding the maximum uncertainty.

For the case where the δ_i are all equal to 1.0 (e.g., where γ_i, s_i, and β_i are all 1.0), then S_δ^2 is 0.0. These conditions correspond to Table III calculations where the standard deviation was an increasing function of ρ. Because the $\{\delta_i\}$ depends on the expert's standard deviations, $\{s_i\}$, there is no reason to expect this to occur.

The expression for the LS variance is

$$\sigma^2 = (\underline{\beta}^{t}\Sigma^{-1}\underline{\beta})^{-1}$$

and it reduces, for $\beta_i/s_i\gamma_i$ all equal to one, to

$$\sigma^2 = [1 + (n-1)\rho] / n\ .$$

Substituting various values for ρ and n=5, this expression gives the variances (and standard deviations) as in Table III.

For the case where all γ_i and β_i are 1.0, but the s_i values, as given in Table II, are {1.3, 1.4, 1.4, 0.7, 1.6}, then the variance expression becomes

$$\sigma^2 = \left(\frac{(n-1)}{(1-\rho)} S_\delta^2 + \frac{n}{[1+(n-1)\rho]} \bar{\delta}^2\right)^{-1} .$$

For this case the values of δ_i are {0.8, 0.7, 0.7, 1.4, 0.5}, giving a δ mean of 0.9 and δ standard deviation (S_δ) of 0.3. Using these values for the various correlations, the above expression for the LS variance gives the those standard deviations listed in Table IV. The mathematical development explains how the unexpected pattern in Table IV occurred for the types of cases in that table.

The expressions developed in this subsection help in the understanding of how the correlation structure relates to the LS standard deviation. Earlier in the method descriptions section, some of the relationships of the other parameters to the LS mean and variance were mentioned. The purpose of this report is to go beyond these parametric relationships and to address the issues of comparing this method to a set of actual values and of comparing its performance to that of other methods.

V. THE VALUES USED IN THE STUDY

In the sensitivity studies and simulation designs described below, parameter values are estimated in each of the four methods. To compare the performance of each method, some basis for comparison is required. In this section, a set of "actual" parameter values is established. In the following sections, the abilities of the methods to recover the actual log failure rate are examined over a range of circumstances, called "cases".

The actual log failure rate, θ, is set to -20.0, well within the range of expert-to-expert values (Table II). The actual parameters for the LS method, α, β, γ, and ρ, are given in Table V. The values for α, which corrects for location or shift bias, were chosen to range from -2.0 to 2.0. This range covers four orders of magnitude. The β values also change the expert's means in combination with the α values. Values for β were chosen across a narrower range so that the combined α, β effects would not produce severe changes over many orders of magnitude. The γ values were chosen to be greater than 1.0 for inflating the standard deviations, countering the common underestimation of the variance. Were the decision maker to be omniscient and know these actual values, the LS aggregated opinion would have a mean equal to -20.0 and the perceived standard deviation σ would equal the actual standard deviation.

TABLE V
ACTUAL VALUES MIXED VALUES FOR α, β, and γ

			Expert		
Parameter	A	B	C	D	E
α	-2.0	0.0	2.0	-1.0	1.0
β	0.925	1.4	1.3	1.1	1.0
γ	2.0	1.5	1.75	2.5	1.25
m	-20.5	-28.0	-24.0	-23.0	-19.0
s	1.3	1.4	1.4	0.7	1.6

where

$$m_i = \beta_i * \theta + \alpha_i,$$

and $\rho_{ij} = 0$ for all $i \neq j$.

The elements of the actual variance-covariance matrix, Σ^a, are $[\sigma_{ij}] = \rho_{ij}\, \gamma_i\, \gamma_j\, s_i\, s_j$. Using the actual Σ, the actual β vector, and the formula for the the variance

$$(\sigma^a)^2 = \frac{1}{(\beta^a)'(\Sigma^a)^{-1}\beta^a} ,$$

the best case perceived mean has a standard deviation, σ^a, of 0.81.

In the application of the LS method, the decision maker estimates the four sets of parameters prior to receiving the experts' mean and standard deviations. His values for these parameters could be the the same across all experts. A decision maker who is not well informed about differences between his experts may be prone to this. The actual values could also be equal across the experts, but experience (Booker and Meyer, 1988) indicates that this is very unlikely. Therefore, some of the sensitivity study calculations (runs) were made using identical values (decision maker's estimates) for α, β, and γ, across all experts. For these runs, comparisons to actual parameter values that are equal across experts were made using the values in Table VI below.

TABLE VI
EQUAL ACTUAL VALUES FOR α, β, and γ

	Expert				
Parameter	1	2	3	4	5
α	1.0	1.0	1.0	1.0	1.0
β	1.2	1.2	1.2	1.2	1.2
γ	1.5	1.5	1.5	1.5	1.5
m	-23.0	-23.0	-23.0	-23.0	-23.0
s	1.3	1.4	1.4	0.7	1.6

and $\rho_{ij} = 0$ for all $i \neq j$.

The values for α, β, and γ were chosen from values in Table V. Specifically, the values are in the middle of the ranges listed in Table V. They should, therefore, be of the order that a decision maker might choose.

VI. THE SENSITIVITY STUDY

The main purpose of the sensitivity study is to quantify the effects that an imperfect decision maker has on the aggregation process. Three different methods (LS, SW, and EW) are directly compared in this study by examining the effects of the decision maker's parameter estimates on the final aggregation mean and variance results, compared to the actual values established.

The Lindley and Singpurwalla Method

The first sensitivity study focused on the features of the LS method. The basic idea behind the decision maker estimating all the parameters is so that he can adjust the experts' mean and standard deviation estimates in an attempt to estimate the actual means and standard deviations. This idea translates to the ideal situation where the decision maker's estimates of the parameters transforms the experts' estimates (means and standard deviations) into the aggregated mean and its variance. The runs were set up such that when the decision maker specified the proper parameters, the resulting value of the aggregation mean estimate, μ, and the perceived variance would be the specified actual values, θ (=-20.0), and the corresponding variance (=0.66).

Another way of looking at this ideal case is by recognizing that the estimates provided by the experts, m, are a function of θ, the actual α and β parameters, and some random error deviation, e,

$$m = \alpha^a + \beta^a \theta + e \ .$$

This error is distributed normally with mean 0 and with a variance/covariance matrix whose elements, $\Sigma_{ij}{}^a = \rho_{ij}{}^a \gamma_i{}^a \gamma_j{}^a s_i s_j$, are based on the actual values. Recall that the LS aggregation mean has the following form

$$\mu = \frac{\underline{\beta}^t \Sigma^{-1} (\underline{m} - \underline{\alpha})}{\underline{\beta}^t \Sigma^{-1} \underline{\beta}} \quad .$$

Rather than designing a simulation constructing expert estimates based on the above formulas with the random error component, the aggregation mean and variance can be found using a normal distribution that is not conditioned on this random error component. This unconditional distribution based on the Theorem 1 (Lindley and Singpurwalla, 1982), gives the aggregation mean estimate, μ, distributed normally with mean

$$\frac{\underline{\beta}^t \Sigma^{-1} (\underline{\alpha}^a - \underline{\alpha} + \underline{\beta}^a \theta)}{\underline{\beta}^t \Sigma^{-1} \underline{\beta}}$$

and variance

$$\frac{\underline{\beta}^t \Sigma^{-1} \Sigma^a \Sigma^{-1} \underline{\beta}}{\left(\underline{\beta}^t \Sigma^{-1} \underline{\beta}\right)^2}$$

where the superscript "a" denotes the actual vector or values of the parameters.

For the ideal case, when the decision maker estimates $\underline{\alpha} = \underline{\alpha}^a$, $\underline{\beta} = \underline{\beta}^a$, $\gamma = \gamma^a$ and $[\rho_{ij}] = [\rho_{ij}{}^a]$, the aggregation mean reduces to the value of θ and the variance of that mean reduces to the familiar variance expression of $1/(\underline{\beta}^t \Sigma^{-1} \underline{\beta})$ from Theorem 1 (Lindley and Singpurwalla, 1986). For nonideal cases, the aggregated mean and standard deviation can differ substantially from actual values.

The actual values for the parameters used in the sensitivity studies are established using the values from the example in tables I and II as a guide. The values for the decision maker's estimates of the parameters are given in the section below describing the design for the sensitivity study runs.

The Self Weight Method

To compare the results of the other methods to the LS method, the unconditional distribution can be structured to resemble the self-weight aggregation method used by

Lawrence Livermore National Laboratory. Therefore, some of the LS runs made using the unconditional distribution with the various combinations of the LS parameter estimates also provide an aggregation mean that has the self-weight form

$$\mu = \sum_{i} w_i m_i$$

where the w_i are normalized (weights sum to 1.0) self weights provided by the experts and m_i are the estimates provided by the experts. Although the original report did not specify one, the variance of this weighted mean could be estimated by the standard formula for the variance of a weighted mean with normalized self weights, w_s,

$$\sigma^2(\mu) = w_s' \Sigma(m) w_s \, ,$$

where $\Sigma(m)$ is the variance matrix of the mean, having the form of a diagonal matrix with $\{s_i^2\}$ elements. Here $\Sigma(m)$ is a general notation for variance to distinguish it from the LS variance, Σ.

In the LS method if the decision maker sets parameter $\underline{\alpha} = \underline{0}$, then the unconditional aggregation mean estimator has the form of the weighted mean estimator where the weights are

$$w_{si} = \frac{\sum_{j} \beta_j \sigma^{ij}}{\sum_{i,j} \beta_i \sigma^{ij} \beta_j}$$

where σ^{ij} are the elements of Σ^{-1}. The variance for the aggregation mean is the unconditional variance formula.

Of the LS cases, the ones where there is no bias correspond to the weighted mean cases. Therefore, both methods can be directly compared using these common cases.

The Weighted Mean Method

The weighted mean method originally proposed by EPRI, was later modified in a second report (McGuire, et al., 1989) such that all the experts were given equal weights. Therefore, this modification was chosen for use in the study and comparison of methods, making the EPRI method a representative of the equal weights method for aggregating multiple estimates.

The use of equal weights is a popular method for aggregating expert estimates (Seaver, 1978; Winkler, 1986). One of the main arguments for choosing equal weights is that the analyst or decision maker doing the aggregating often has no information or no reason for choosing the weight of one expert over another. Another reason stems from the fact that minor changes in weights have little impact on the resulting weighted mean; however it can affect the perceived variance (Meyer and Booker, 1991).

In the cases run using the unconditional distribution, only one combination of the LS parameters corresponds to the equal weights case. Thus the results for this method are limited to only that one case. However, the comparison among the three methods is still feasible even if there is only one case for the equal weights method.

That one case is where all $\alpha = 0$, $\beta = 1$, $\rho_{ij} = 0$ for $i \neq j$ and $\gamma_i = 1/s_i$, making $\Sigma = I$ such that the weights in the aggregation mean are equal. This case corresponds to the situation where the decision maker makes no location or scale adjustments to the experts' means, and he happens to select γ_i values that rescale their stated variances to 1.0. These parameter choices are not very likely nor do they have a reasonable foundation except to achieve the goal of weighting the experts equally.

VII. THE SIMULATIONS

The Empirical Distribution Method Simulation

Using the empirical distributions of the experts' estimates from Table I, a simulation was designed to examine the use of this method. This simulation does not have a direct connection to the other three methods nor to the unconditional sensitivity studies comparing them.

The empirical distributions of the experts are based on a set of five estimates from the experts: an absolute minimum estimate (cumulative probability of 0.0), an absolute maximum estimate (cumulative probability of 1.0), a middle value (best guess), an upper tail value, and a lower tail value. There is an ongoing debate about what percentiles should correspond to the two tail estimates and middle estimates (Meyer & Booker, 1991). Some suggest that the expert's best or middle estimate represents a mean value; others suggest that it represents the median (50th percentile). Some suggest that the tail values represent the 5th and 95th percentiles while some studies suggest that the tail values represent the 40th and 60th percentiles. The simulation is designed to investigate the effects of changing these percentile levels for the tail and middle values. In other words, the simulation examines how misspecification of the percentile levels affect the results.

The Self Weight Method Simulation

Another simulation was designed to investigate the effects of changing the experts' weights that was not tied to the LS method. In the original application of this method, the weights were integer values from 1 to 10. These values were then normalized so that the experts' weights summed to 1.0.

In this simulation, the weight values (1 to 10) were simply generated from two different probability distributions. The first was a skewed distribution (Figure 2) which placed larger probabilities on the higher ratings (6-9) and very little probability on the lower ratings (1-4). The reason for this comes from the premise that most experts should be able to rate themselves according to their expertise, which should be on the high side of the scale. However, there many other factors to consider concerning self weights such as the expert's own degree of confidence, his actual level of expertise versus his perceived level, and his expertise relative to others. All of these factors make it difficult to specify a distribution for the weights, especially with no guidance from the literature. To provide an alternative, the second probability distribution was a uniform on the integers 1 to 10 (Figure 3).

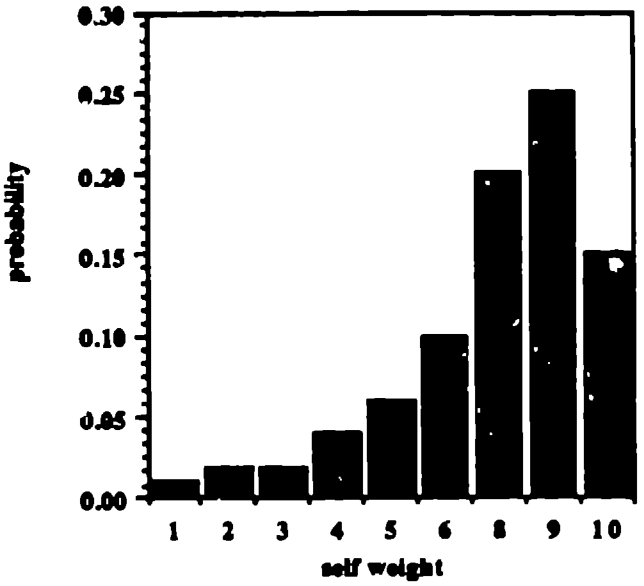

Figure 2: Skewed Distribution for Determining Self Weights

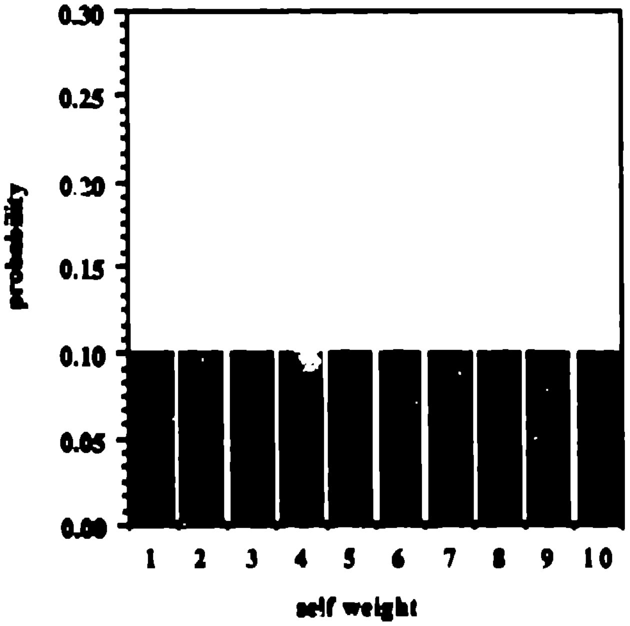

Figure 3: Uniform Distribution for Determining Self Weights

VII. DESIGN OF THE SENSITIVITY & SIMULATION RUNS

There were five different sensitivity study and simulation runs made using three different FORTRAN codes. Using the first of these codes, the equal weight, self weight and LS methods were compared using the same cases and the same actual values. The cases and sets of actual values formed three different types or runs: (1) equal actual values (Table VI) compared to the decision maker's equal estimates of the parameters, (2) mixed actual values (Table V) compared to the decision maker's equal estimates of the parameters, and (3) mixed actual values (Table V) compared to the decision maker's estimates based on small deviations from the actual values. Using the second FORTRAN code, the self weights for that method were generated from both the uniform and skewed distributions. Using the third FORTRAN code, the empirical cumulative distribution functions of the experts were used to form an aggregate mean and median estimators. These five different runs are further described below.

Run 1 - Equal Values

There were 301 cases run involving the EW, SW, and LS methods. The parameter values used are in Table VI, and the cases are constructed from combinations of the parameters in Table VII (the 300 cases equal the possible combinations of five α values, five β values, four γ values, and three ρ matrices) plus the single case for equal weights. Here the actual values for the parameters are constant across experts, and the decision maker estimates the same values for all the experts (Table VII).

TABLE VII
THE DECISION MAKER'S EQUAL ESTIMATES FOR α, β, γ, and ρ

	Parameter			
	α	β	γ	ρ
LS Cases	-2.0	0.8	1.0	0.0
	-1.0	1.0	1.5	0.5
	0.0	1.2	2.0	0.9
	1.0	1.4	2.5	
	2.0	1.5		
Equal weight	0.0	*	1.0	0.0

*={0.78, 0.71, 0.74, 1.43, 0.64}

The case of $\alpha = 1.0$, $\beta = 1.2$, $\gamma = 1.5$, and $\rho = 0.0$ corresponds to parameter choices by a perfect decision maker. All the other cases correspond to parameter choices of an imperfect decision maker

Run 2 - Unequal Values

There were 301 cases run involving the EW, SW, and LS methods. The actual values for the parameters are in Table V, and the cases are constructed from all possible combinations of estimates in Table VIII (300 cases) plus the single case for equal weights. Here the actual values for the parameters are mixed values, and the decision maker's estimates the same values for all the experts (Table VIII).

TABLE VIII
THE DECISION MAKER'S EQUAL ESTIMATES FOR α, β, γ, and ρ

	Parameter			
	α	β	γ	ρ
LS Cases	-2.0	0.8	1.0	0.0
	-1.0	1.0	1.5	0.5
	0.0	1.2	2.0	0.9
	1.0	1.4	2.5	
	2.0		3.0	
Equal weight	0.0	*	1.0	0.0

*={0.78, 0.71, 0.74, 1.43, 0.64}

Run 3 - Small Deviations

There were 300 cases run involving the EW, SW, and LS methods. The actual values for the parameters are in Table V, and the cases are constructed from the combinations of the parameters in Table IX. Here the experts' actual values for the parameters are mixed values, and the decision maker estimates the same values for the experts that are almost identical to the actual values (Table IX). In each case, the estimate for only one parameter, for only one expert differs from the actual values. Here, the decision maker is nearly perfect in his estimation process – an unrealistic expectation.

TABLE IX
THE DECISION MAKER'S SMALL DEVIATIONS FROM THE ACTUAL MIXED VALUES FOR α, β, γ, and ρ

			Expert		
Parameter	A	B	C	D	E
α Cases	-1.0	0.0	2.0	-1.0	1.0
	-2.0	1.0	2.0	-1.0	1.0
	-2.0	0.0	1.0	-1.0	1.0
	-2.0	0.0	2.0	0.0	1.0
	-2.0	0.0	2.0	-1.0	0.0
β Cases	1.2	1.4	1.3	1.1	1.0
	.925	1.2	1.3	1.1	1.0
	.925	1.4	1.2	1.1	1.0
	.925	1.4	1.3	1.2	1.0
	.925	1.4	1.3	1.1	1.2
γ Cases	2.0	2.0	1.75	2.5	1.25
	2.0	1.5	2.0	2.5	1.25
	2.0	1.5	1.75	2.0	1.25
	2.0	1.5	1.75	2.5	2.0

ρ_{ij} Cases ρ_{ij} = {0 .0, or 0.5, or 0.9} for all i≠j.

Run 4 - the Distribution Self Weights Simulation

There were 1000 simulation samples run on the SW method where the weights were generated from a skewed-shaped distribution (Figure 2) and 1000 samples from a uniform distribution (Figure 3). The structure of this simulation did not use the LS framework. The weights were randomly chosen values from the 1 to 10 scale according to the two distributions. The weights were then normalized as in the original method. The weighted aggregation mean and its standard deviation were calculated and compared to the actual mean and variance.

Run 5 - the Empirical Distribution Simulation

There were 300 cases run on the empirical distribution method. These cases were constructed using 300 different sets of expert distributions. The distributions for each expert were formed using the upper tail, lower tail, maxima, minima and central values given in Table X. The tail and extreme values are based on the experts' estimates given in Table I, and the central values are based on the experts' estimates given in Table II.

TABLE X
VALUES FOR SPECIFYING THE EXPERTS' EMPIRICAL DISTRIBUTIONS

			Expert		
Values	A	B	C	D	E
minimum	-23.9	-32.6	-20.0	-29.8	-24.0
lower tail	-22.3	-30.3	-25.2	-25.9	-22.8
central	-20.5	-28.0	-24.0	-23.0	-19.0
upper tail	-19.1	-23.7	-20.6	-20.6	-16.6
maximum	-17.5	-23.4	-17.2	-16.9	-14.0

The empirical distributions were constructed by randomly assigning percentile levels to the two tail and the central values. The maxima and minima values anchor the distributions at probabilities 1.0 and 0.0, respectively. Linearly interpolating the five points forms the empirical distributions. The percentile level for the lower tail values for each expert were randomly assigned according to a uniform distribution on the range 0.0 to 0.30, the percentile level for the upper tail values from a uniform ranging from 0.70 to 1.0, and the percentile level for the central value from a uniform ranging from 0.40 to 0.60. The ranges of the uniforms were chosen to cover the ranges suggested by various studies (Kahneman, et al., 1982). A single case is the set of five expert distributions that were formed by the percentile assignments made at random for the two tail and central values.

For each of the 300 cases, the five expert's distributions are aggregated forming a final combined distribution whose (unweighted) mean, median, variance and percentiles can be used as summary statistics. To form this final distribution, values are sampled from the five expert's distributions. These values are combined according to one of two chosen aggregation estimators, the mean and the median. In other words, a sample is a set of five randomly selected values, one from each expert distribution. The mean and the median of the five values is calculated. This sampling was done 400 times producing a final distribution for the mean and a final distribution for the median. The values of the two aggregation estimators (the mean and median) were taken as the central values of these two final distributions. Likewise, standard deviations were calculated from these two final distributions. Probability intervals (not confidence intervals) for the perceived mean were calculated using the percentiles from the final distributions.

Measures of Comparison

All the sensitivity studies and simulations used three different measures of comparison to their respective actual value sets. The first two measures are simple deviations: the aggregation estimate for the method from the value of θ and the estimator's standard deviation estimate from the actual standard deviation

$$\Delta\mu = \theta - \mu \quad \text{and} \quad \Delta\sigma = \sigma^{a} - \sigma \ .$$

The value of σ varies from case to case. This comparison is based on the difference between the case-specific perceived standard deviation and the case-specific actual standard deviation.

The third measure is based upon how well each method covered the value of θ using probability or confidence intervals calculated from the estimates of the aggregation mean and its standard deviation

$$P\{\theta \in \mu \pm z\sigma\} -[\Phi(z) - \Phi(-z)]$$

where z a percentile of a Normal (0,1) and Φ(z) is the corresponding cumulative distribution function. For example, when a 90% interval is calculated based upon the estimates for a given method, in what percentage of the cases run does that intervai cover θ? The expected answer should be near 90%. For each method, five different sized intervals were calculated for the 50%, 65%, 80%, 90% and 99% coverages. The results are given in the section below.

VIII. RESULTS

Comparison to the Actual Values for all Methods

The measures of comparison described in the previous section produced the results in Tables XI and XII. The values in Table XI are averages over all the cases. The exception being that the EW values represent only one case.

TABLE XI
DEVIATIONS FROM THE ACTUAL MEAN (θ) AND STANDARD DEVIATION (σ^a) FOR THE MEANS AND STANDARD DEVIATIONS OF THE METHODS.

method/run	mean	deviation from θ	standard deviation	deviation from σ^t
LS / 1	-20.1	-0.1	1.1	0.5
LS / 2	-20.3	-0.3	1.7	0.9
LS / 3	-20.0	0.0	1.5	0.7
SW / 1	-23.0	-3.0	0.7	0.1
SW / 2	-23.1	-3.1	1.0	0.2
SW / 3	-23.3	-3.3	0.8	0.0

TABLE XI (continued)

method/run	mean	deviation from θ	standard deviation	deviation from σ^t
EW / 1	-23.0	-3.0	1.3	0.7
EW / 2	-22.9	-2.9	3.7	2.9
SW /4 skw	-22.5	-2.5	0.6	-0.2
SW /4 unif	-22.5	-2.5	0.6	-0.1
ED / 5 mean	-22.6	-2.6	1.5	0.7
ED / 5 median	-22.3	-2.3	1.8	1.0

where the numbers (1-5) refer to the type of run made;

mean is the average of the 5 experts was used as the aggregation estimator;
median is the median of the 5 experts was used as the aggregation estimator;
skw is self weights generated from a skewed distribution; and
unif is self weights generated from a uniform distribution.

TABLE XII
PERCENT COVERAGES OF θ BY 50, 65, 80, 95 AND 90 PERCENT INTERVALS

method/run	50%	65%	80%	90%	99%
LS / 1	9.6	15.3	21.3	27.6	41.5
LS / 2	10.0	15.0	20.7	27.8	41.0
LS / 3	53.0	67.0	73.0	83.0	92.3
SW / 1	0.0	0.0	13.2	21.2	53.9
SW / 2	0.0	0.0	13.3	20.0	46.7
SW / 3	0.0	0.0	0.0	0.0	50.0
SW / 4 skw	0.1	0.2	0.4	0.8	5.1
SW / 4 unif	2.2	3.6	5.5	8.7	20.7
ED / 5 mean	0.0	0.0	0.0	0.0	45.0
ED / 5 median	0.0	0.0	9.3	58.7	98.3

The averaging process in Table XI requires careful interpretation, especially for the aggregation estimators (means). For example, the LS run 1 has a very small overall deviation from θ; however, variation in the aggregation means over the 300 cases is quite large. The highest case was -10.0 while the lowest case was 31.0. The low cases occurred when α was the largest (2.0) (reducing μ) and when β was the smallest (0.8) (also reducing μ) for the various values of γ and ρ. The high cases occurred when the correlation was the highest (0.9), α was the smallest (-2.0), and β made no adjustments (1.0) for various values of γ. The SW run 1 cases all have the same aggregation estimator (= -23.0). This results from the fact that the SW cases are restricted to combinations of α, β, γ, and ρ such that the resulting weights are normalized. The 75 cases for the SW method, where the weights are automatically normalized, correspond to LS cases where $\rho = 0$ and $\beta \geq 1$. The EW case is also restricted and misses the target mean. The standard deviation, although somewhat larger, is not as large as might be expected. Recall that it is a composite of two sources of variation that one might suspect would produce a large value.

For run 2, again the LS overall mean is on target with θ. However, again there is a very large variation in these mean values ranging from -31 to -16. The 75 cases in the SW method have means with a much smaller variance ranging from -22.9 to -23.3. These 75 cases again correspond to those where $\rho = 0$ and $\beta \geq 1$. The EW case here does have a very large standard deviation as might be expected from the way it is calculated.

By construction, the small deviations cases (run 3) for the LS method have means that are much less variable, ranging from -23.8 to -16.7, and their overall average is θ. Although the variation in the 110 means applicable to the SW method is also this small, the overall mean for these cases is not on target with θ. These 110 cases are defined from combinations of $\alpha = -1$ or -2, $\beta = 0.925$ or 1.2, $\gamma = 2.0$ and $\rho = 0.0$ or 0.5. The EW method is not applicable here.

In run 4, where SW method weights are generated from two different distributions instead of being defined in terms of the LS cases, the overall means are not on target with θ. The overall means from the two different distributions are nearly identical, indicating little effect from the different distribution (skewed versus uniform) used for the weights. (This is not surprising because for any distribution with independent, identically distributed weights, the expected value of the normalized weights is 1/n. The result is equal weights on the average. Only the variance should be expected to change.) The variation in the means for the 1000 cases is also small, ranging from -25.6 to -19.4 for the uniform distribution and from -24.5 to -20.3 for the skewed distribution.

The means and medians from the 300 ED cases in run 5 both differed from the target θ. The mean values ranged from -23 to -22 and the medians from -22.9 to -21.5, indicating very small variation for both types of estimators for run 5 cases versus the other runs.

It is not surprising that the overall means for many of the runs missed the target value. The EW method and run 5 weight the five experts equally. This weighting is not conducive to hitting the target value because -20.0 is not in the center of the five experts' estimates or their distributions. (This is not an unusual occurrence because it would be unreasonable to expect the experts' estimates or distributions to be exactly centered around the actual value.) In addition, when there is small variation for the aggregation estimator, such as in the ED method and the SW method, it is difficult to hit the target value. The methods that can consistently hit the target value, such as the LS method, are those which either have large variability about the aggregation estimator.

Making comparisons of the overall standard deviations is not as clear cut as comparing to the target θ because σ^a is based on the LS variation. Even so, the LS standard deviation of the aggregation mean compares poorly to the target value, σ^a. The ED method produces standard deviations that are about twice σ^a. The only method that consistently hits the desired standard deviation for the various runs is the SW method.

The third measure of comparison examines how well the confidence and probability intervals cover the target mean. These results are in Table XII and are graphically represented in Figure 4 (except for the EW method).

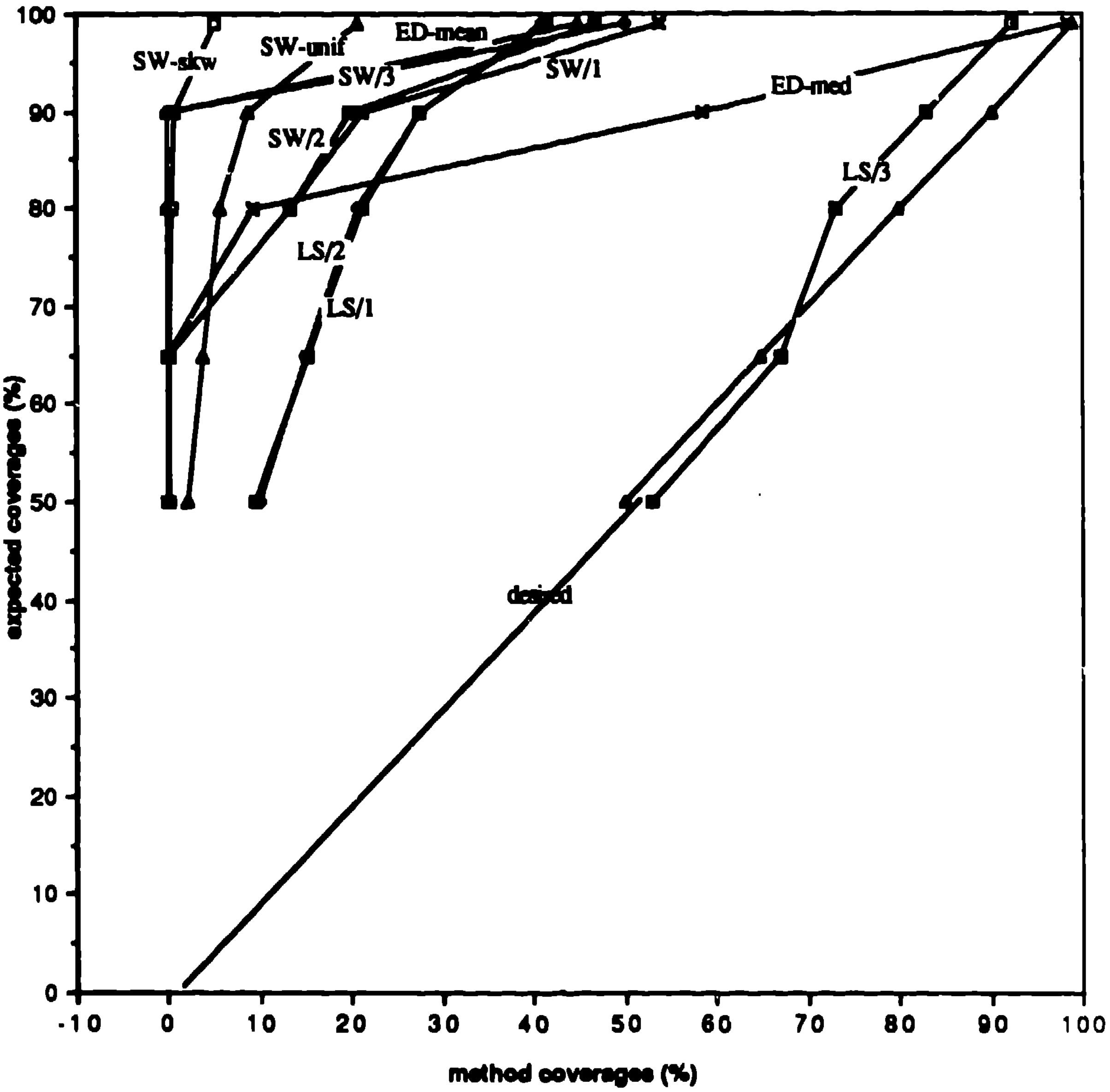

Figure 4: Target Mean Coverages Based on Various Sized Intervals

The coverages that one would hope to achieve should be near the "desired" line. Ideally, a 50% interval should cover the target value of -20.0 in 50% of the cases run. However, a decision maker using imperfect parameter values can be seriously misled as to coverage rates. Intervals having perceived coverages of 95% often had actual coverage of less than 50% in this study

Even though the LS mean was very near -20.0 (Table XI), that result was obtained by off-target high intervals cancelling off-target low intervals. Low mean values for other methods were attributable to expert opinions being consistently pessimistic.

The results from Table XII and Figure 4 indicate that many of the methods did rather poorly in achieving the expected coverages. Based on the results already mentioned above regarding the target misses and the variability associated with that process, it is not surprising that the intervals for those methods exhibit poor coverage of the target mean. After all, those intervals are calculated from the estimated aggregation mean and its estimated standard deviation, and the expert makes no allowance for the possibility that some of the input parameters could be slightly in error, or that the experts themselves might be imperfect as a group.

The only method and simulation combination that resembles the expected coverage in Figure 4 is the LS method for small deviations. The ED method using the median estimator does better for larger sized intervals (≥ 90%). The other coverages from run 1 and run 2 for LS are nearly identical, as are the coverages for run 1 and 2 for SW. The poorest coverages are for the SW method where the weights are generated from the uniform and skewed distributions.

Comparison to the Actual Values for Correlation Effects

The results for runs 1-3 for the LS method include three different correlation structures. The actual values structure has off diagonal elements of the correlation matrix equal to 0.0, and the cases run include this structure plus structures where these elements are equal to 0.5 and 0.9.

Examining the measures of comparison for the three different correlation structures gives the following results. For runs 1 and 2, the means and standard deviations over the 100 cases for each correlation structure do not change as the structure changes. The percent coverages for the various interval sizes (50%, 65%, 80%, 90%, and 99%) also do not change with the different correlation structures.

For run 3, the means do not change, but the standard deviations increase from 0.82 for 0.0 correlation to 2.63 for correlations of 0.9. Also for run 3, the coverages for the various interval sizes also change as indicated in Table XIII. For each interval, the coverage for the correlation of 0.5 is approximately that listed in Table XII. However, the coverage for 0.0 correlation is much higher, quickly approaching 100%. The coverage for correlation of 0.9 is much lower, approaching the coverages seen in the other methods. Therefore, when the decision maker estimates high correlation among experts (≈ 0.9) when, in truth, none exists, poor coverages result even when he has correctly estimated all parameters of all experts except one. The extreme sensitivity of target coverage is indicated in these results.

TABLE XIII
PERCENT COVERAGES OF θ BY 50, 65, 80, 95 &90 PERCENT INTERVALS
FOR DIFFERENT CORRELATIONS IN THE LS METHOD, RUN 3

decision maker's correlation*	50%	65%	80%	90%	99%
0.0	84	97	100	100	100
0.5	52	67	78	92	100
0.9	23	37	41	57	77

*Actual correlation is 0.0.

IX. CONCLUSIONS

Conclusions from the results of the three measures of comparison are mixed and do not indicate any one method as better than the others. All methods are extremely sensitive to imperfect estimations of the parameters whether these inputs come from the experts themselves or the decision maker. Coverage rates for all methods are also poor. The only exception is when the inputs are nearly perfectly estimated, and this is unrealistic to expect.

Some interesting conclusions about the methods were found based on the features of the methods. Four different features are worth mentioning: how the method handles correlation, how the method handles uncertainty, what assumptions are required for use, and what input estimations are required for use.

The LS method is the only one that specifically addresses the correlation structure among the experts. The difficulty here is that the decision maker, analyst, or whoever is aggregating the information must estimate that correlation structure. Studies are lacking on how to do this estimation either in controlled or realistic environments. In the LS runs, there is some evidence for differences in the results depending upon which of the three correlation structures were specified.

Uncertainty in the experts' estimates is handled in the methods by two ways. The variance estimates elicited for use in the LS method is one way. The percentile estimates elicited and corresponding empirical distributions in the ED method is another way. The two weighted methods (EW and SW) do not suggest eliciting uncertainty estimates from the experts in either form. This is a shortcoming of these methods, as situations requiring only a point estimate are rare. Uncertainty is present in the elicited data, and this uncertainty should be represented in some way.

Problems arise in using methods that rely on restrictive assumptions on the data. In general problems arise from certain properties in the data. For example, there is a basic problem with using any statistical technique for analysis that requires the data to be independently gathered, identically distributed sample points from an underlying population. It can be argued that expert judgments may not be an independent or identically distributed sample from a population. However, this independence applies to the method of sampling and does not pertain to the observed values. As a consequence, independence-based inference about the population from the judgments may not apply. Another example, the common use of the central limit theorem (which states that the sample mean is asymptotically normal) may not be appropriate because the underlying population can be highly skewed and the sample sizes are extremely small. In addition to general requirements, some methods (e.g. the LS method) have assumptions, such as the mean must be normally distributed, which may not apply. As a result, methods that require fewer assumptions about the data (e.g., simulation-based, data-based, and non-parametric methods) are more desirable for use in expert judgment analysis when the experts are not "distorted" relative to reality. For example, the ED method uses data-based distributions of the experts, and aggregates them using a simulation-based technique. The SW and EW methods do not require distributional assumptions on the data; however, the variance estimates of their aggregated means do rely on the same sampling asymptotic theory that the LS method does.

All the methods require the experts to make estimates. Two require them to estimate some measure of uncertainty/variability as well. However, the LS method requires the decision maker or analyst to estimate three additional parameter vectors and one additional matrix for the experts. The SW method also requires the experts to estimate their own weights. Some studies have indicated that experts are not good at estimating uncertainty (Kahneman, et al., 1982), and some have suggested that experts are not good at estimating self weights (or the weights of others) (Bernreuter, et al., 1989). There is little evidence to predict how well the decision maker can estimate the parameters needed for the LS method. One can only speculate that the decision maker's estimation of these parameters would not be any more accurate than his ability to estimate variability (which studies have indicated is not done well). Also, the decision maker is presumably less knowledgeable than his experts which binds his estimation ability. Therefore, any method that relies heavily on estimates (other than the estimates of the target quantity) is introducing additional sources of potentially large variability. As a consequence, a "simpler is better" philosophy has developed. This philosophy has led to the conclusion that equal weights for all experts might be the answer. However, the results of this study do not necessarily support the use of equal weights. While keeping additional estimations to a minimum is reasonable advice, for analysis purposes it is better to have estimates of uncertainty on hand.

Based on these features of the methods, the conclusion is that all the methods have their advantages (ease of use, characterizing uncertainty, and accounting for correlation) and disadvantages (restrictive assumptions, imperfect input estimation, and poor performance in the cases run) for use in aggregating expert judgments. The following recommendations summarize the results of this study, focusing on the balance between these advantages and disadvantages. If enough information is known about the experts, and if the target quantity of interest (failure rate) is thought to be normally or lognormally distributed, the LS method can be used with decent results. If both of these conditions are not satisfied, then an alternative method is the ED using the median aggregation estimator.

At the beginning of this study, we did not expect any one of the chosen methods to emerge as the best. If a proven best, method existed, it would probably have been publicized by now. The problems encountered in analyzing expert judgments are numerous, complicated, and frustrating. Research and practical use of proposed analysis methods is sparse. The data itself lacks properties necessary to support both statistical and cognitive theory. Often the theories conflict, e.g. the data are distributed normally for the analysis, but cognitive studies indicate non-normality. As a consequence, guidance for handling the aggregation problem is given primarily from the theoretical side and not from the experimental or the experience sides. In reality, there is a decision maker or analyst who is faced with providing aggregation estimates for documentation or justification purposes.

This study focused on a real example and the use and comparison of four different types of methods that have received attention in the literature. This study provides insights into these methods, and the decision maker or analyst can gain some guidance from the results.

ACKNOWLEDGMENT: This work was performed at the Statistics Group at Los Alamos National Laboratory and was sponsored by the Nuclear Regulatory Commission, Office of Nuclear Regulatory Research , FIN A7225-0, under the direction of Dr. James W. Johnson.

REFERENCES

Booker, J. M. and Meyer, M. A. (1988), "Sources and Effects of Interexpert Correlation: An Empirical Study," *IEEE Transactions on Systems, Man and Cybernetics*, 18, 135-142.

Bernreuter, D. L., Savy, J. B., Mensing, R. W. and Chen, J. C.(1989), "Seismic Hazard Characterization of 69 Nuclear Plant Sites East of the Rocky Mountains", NUREG/CR-5250, UCID-21517, vol. 1.

Genest , C. and Zidek, J. V. (1986). "Combining Probability Distributions: A Critique and an Annotated Bibliography", *Statistical Science*, vol. 1, p. 114-148.

Kahneman, D., Slovic, P. and Tversky, A. Eds. (1982). *Judgment Under Uncertainty: Heuristics and Biases*. Cambridge, Massachusetts: Cambridge University Press.

Lindley, D. V. and Singpurwalla, N. D. (1986), "Reliability (and Fault Tree) Analysis Using Expert Opinions", *Journal of the American Statistical Association*, vol. 81, no. 393, p.87-90.

McCann, M. W., Jr. (project manager) et al. (1988), "Seismic Hazard Methodology for the Central and Eastern United States," EPRI report NP-4726-A, vol. I, parts 1 & 2.

McGuire, R. K. (principal investigator), et al. (1989), "Probabilistic Seismic Hazard Evaluations at Nuclear Sites in the Central and Eastern United States: Resolution of the Charleston Earthquake Issue," EPRI report NP-6395-D.

Mensing, R. W. (1990) Private communication, November 13, 1990.

Meyer, M. A. and Booker, J. M. (1991), *Eliciting and Analyzing Expert Judgment: A Practical Guide*, Academic Press: London.

Seaver, D. A. (1978) "Assessing Probability with Multiple Individuals: Group Interaction Versus Mathematical Aggregation," SSRI Research Report 78-3, Social Science Research Institute, University of Southern California, Los Angeles, California.

U.S. Nuclear Regulatory Commission (NRC), Office of Nuclear Regulatory Research (1989). "Severe Accident Risks: An Assessment for Five U.S. Nuclear Power Plants, vol. 1-2, second draft for peer review, Nuclear Regulatory Commission Report NUREG-1150, Washington, D.C.

Wheeler, T. A., Hora, S. C., Cramond, W. R. and Unwin, S. D. (1986), "Analysis of Core Damage Frequency From Internal Events: Expert Judgment Elicitation," NUREG/CR-4550, SAND86-2084, vol. 2, parts 1 & 2.

Winkler, R. L. (1986) "Expert Resolution," *Management Science*, 32, 298-303.

APPENDIX - NOTATION

α bias parameter vector estimates in the Lindley/Singpurwalla (LS) method

β parameter vector estimates in the LS method scaling the expert's estimate of the mean

γ parameter vector estimates in the LS method scaling the expert's estimate of the standard deviation.

ρ_{ij} estimate of the correlation between experts i and j

Λ the actual value of the failure rate

θ the actual value of the log failure rate (-20.0)

μ the aggregation estimator of θ

Σ estimate of the variance/covariance matrix in the LS method

P the correlation matrix in the LS method

σ standard deviation of the aggregation mean estimator.

m the expert's estimate or best guess of the log failure rate

s the expert's estimate of the standard deviation

σ^a the actual value of the standard deviation of the quantity being estimated

α^a the actual value of the bias parameter in the LS method

γ^a the actual value of the standard deviation scaling parameter in the LS method

β^a the actual value of the scaling parameter for the expert's estimate of the mean in the LS method

ρ^a the actual value of the correlation between experts

Σ^a the actual value of the variance/covariance matrix in the LS method

t indicates transpose of a matrix or vector

_ indicates a vector